DOCUMENT RESUME

ED 128 464 TM 005 657

AUTHOR	Jouett, Michael L.
TITLE	The Internal Validation of Level II and Level III Respiratory Therapy Examinations. Final Report.
INSTITUTION	American Association for Respiratory Therapy, Dallas, Tex.
SPONS AGENCY	Health Resources Administration (DHEW/PHS), Bethesda, Md.
PUB DATE	1 Apr 76
CONTRACT	HRA-231-75-0201
NOTE	46p.
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	*Certification; Criterion Referenced Tests; *Equivalency Tests; *Inhalation Therapists; Item Analysis; Norm Referenced Tests; Skills; Statistical Analysis; Test Construction; *Test Reliability; *Test Validity
IDENTIFIERS	American Association for Respiratory Therapy

ABSTRACT

This project began with the delineation of the roles and functions of respiratory therapy personnel by the American Association for Respiratory Therapy. In Phase II, The Psychological Corporation used this delineation to develop six proficiency examinations, three at each of two levels. One exam at each level was designated for the purpose of the validation process. Statistical analysis included the means and standard deviation of the two tests, correlation of scores between these tests and the Certification Examinations and Written Registry Examinations, and an item analysis of the two tests. In retrospect, the original delineation of roles did not provide sufficient behavioral specificity for the derivation of criterion referenced examinations. At this time, a project to "define" respiratory therapy competence is in development to supplant the original delineation of roles document. This definition will then be used as a basis for developing at least one evaluative simulation as a possible alternative assessment form within the credentialing system. (Author/BW)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED128464

Final Report
Contract HRA 231-75-0201

THE INTERNAL VALIDATION OF LEVEL II AND LEVEL III RESPIRATORY THERAPY EXAMINATIONS

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

April 1, 1976

Submitted by
Michael L. Jouett, ARRT
Director of Education
AMERICAN ASSOCIATION FOR RESPIRATORY THERAPY
7411 Hines Place
Dallas, Texas 75235

TM005 657

TABLE OF CONTENTS

Project Overview 2
Methodology. 3
Results. 5
Discussion 17
Recommendations. 22
References 23

APPENDICES

A - Letters to Candidates
B - Biographical Data Sheet
C - Letters of Agreement with ETS
and Psychological Corporation
D - CRM Methodology

PROJECT OVERVIEW

INTRODUCTION

The data-base for the development of Proficiency Examinations came from a Phase I effort by the American Association for Respiratory therapy in its Final Report for Contract 72-4219, DELINEATION OF ROLES AND FUNCTIONS OF RESPIRATORY THERAPY PERSONNEL. The DELINEATION set forth minimal competencies in terms of skills and knowledge necessary to perform respiratory therapy services safely and effectively.

In Phase II, The Psychological Corporation used the DELINEATION to develop six Proficiency Examinations, three at each of the two levels. One exam at each level was designated for the purpose of the validation process. It was this form (03) of the examinations that were used in the internal validation project.

METHODOLOGY

In accord with the recommendations of the Advisory Committee appointed during Phase I of the project, Form 03 of each of the Proficiency Examinations was administered in conjunction with the appropriate credentialing examination of The National Board for Respiratory Therapy, Inc. (NBRT). The Level II exam was given by The Psychological Corporation as Part II of the Certification Examination, while Educational Testing Service administered the Level III exam as Part II of the Registry Written Examination. In order to inform candidates about the combining of the Proficiency Examinations with the regularly scheduled credentialing exams, a letter (See Appendix A) was sent to each candidate by the chairman of the appropriate examination committee. Although examinees were not specifically advised that Part II was an experimental examination, they were aware that results of this section could not decrease their chances of a credential; only increase the raw score on Part I.

A biographical data sheet was designed (See Appendix B) to determine the nature and depth of experience of examinees. Identifying information about each of the candidates was coded to enable studies of possible relationships between competence in respiratory therapy and type or quantity of training and experience. Significantly, approximately 75 registered (ARRT) respiratory therapists took the Level III exam and approximately 50 certified (CRTT) technicians took the Level II examination.

Due to a prior commitment with ETS for the development and annual administration of the Registry Written Examination, it was necessary for NBRT to further utilize their services for the administration of the Level III Proficiency Exam. Consequently, it was necessary to establish a cooperative mechanism among the concerned parties (AART, NBRT, ETS, and Psychological Corporation) in order to successfully accomplish the required test administration and data analysis procedures. Although a satisfactory arrangement was negotiated, results of the Registry Exam were not provided to the Association at the same level of detail as for the other measures. The letters of agreement between AART and ETS and AART and Psychological Corporation may be seen in Appendix C.

On March 15, 1975, Educational Testing Service administered both the NBRT Written Registry Examination and the Level III Proficiency Examination to 2288 registry candidates. Following testing, ETS returned all Proficiency Examination materials to The Psychological Corporation, including test booklets, answer sheets and questionnaires, retaining the registry examination materials for analysis. The Psychological Corporation then computed the Level III total scores and subscores. These data were recorded on tab cards, identifying individuals by name, and returned to ETS.

ETS then matched total scores and subscores for each candidate on the Proficiency and Registry Examinations, recorded the data on

tab cards, deleting identification of individual candidates by name, and returned the cards to Psychological Corporation. After compiling responses to the questionnaires and performing the data analysis, Psychological Corporation forwarded the results to the Association.

On May 10, 1975, Psychological Corporation administered both the NBRT Written Certification Examination and the Level II Proficiency Examination to 3921 certification candidates. All test materials for both exams were retained by The Psychological Corporation, who, after scoring answer sheets, compiling responses to the questionnaires, and analyzing the data, forwarded the results to the Association.

All data analyses were performed under subcontract with The Psychological Corporation. The analyses recommended by The Corporation and specified in Contract HRA 75-0201 were as follows:

1. Means and standard deviation on Level II and Level III
2. Correlation of scores between the Level II and Certification Examinations
3. Correlation of scores between the Level III and the Written Registry Examination
4. Item analysis of the Level II and Level III Examinations, yielding difficulty and discriminatory indices, and showing the point biserial correlation between item performance on the certification of Registry Examination, in addition to that between item performance and performance on Proficiency Examinations.

Although the appropriateness of these analyses is now under question by the Association, the results of all but one of these analyses are presented in the following section according to the contractural agreement. Because the necessary Registry Exam data were not released by NBRT, it was not possible to include correlation of Level III Proficiency Exam items with performance on the Registry Exam.

RESULTS

The raw test data were analyzed by The Psychological Corporation and the results provided to the Association. Statistical characteristics of the Level II and Level III Proficiency Examinations, the Certification Examination for Respiratory Therapy Technicians, and the Written Registration Examination for Respiratory Therapists are presented in Table 1. The reliability and standard error of measurement for the Written Registry Examination were not provided to the Association by ETS and could not be included in the table.

TABLE 1

Summary Statistics for the Level II and Level III Proficiency Examinations and the NBRT Credentialing Examination

	Level II		Level III	
	Prof. Exam	Cert. Exam	Prof. Exam	Reg. Exam
Number of Examinees	3921	3921	2288	2288
Mean Raw Score	171.187	175.350	185.880	128.447
Variance	547.262	792.566	297.578	475.183
Standard Deviation	23.394	28.153	17.250	21.799
Standard Error	7.210	6.991	6.832	N[1]
KR21 Reliability	0.905	0.938	0.843	N[1]
KR20 Reliability	0.926	N[1]	0.886	N[1]

[1]Not provided to the Association

The Proficiency Examinations consist of a series of subtests derived from each of the Level II and Level III categories of the DELINEATION OF ROLES AND FUNCTIONS OF RESPIRATORY THERAPY PERSONNEL (AART, 1973) which were defined during Phase I of the project. Examinee performance in the form of group means and standard deviations is provided for each subscale of the Proficiency Exams in Tables 2 and 3.

TABLE 2

Level II Proficiency Examination Performance by Role and Function (N = 3921)

	No. of Items	Mean	S.D.
Total Score	250	171.2	23.4
Intermittent Positive Pressure Breathing	50	35.5	5.1
Humidity/Aerosol Therapy	38	23.6	4.4
Gas Therapy	50	33.6	6.1
Pulmonary Drainage Procedures	38	26.1	4.0
Cardiopulmonary Resuscitation	12	11.3	2.0
Cardiorespiratory Drug Administration	50	33.4	5.4
Infection Control	12	7.7	1.7

TABLE 3

Level III Proficiency Examination Performance by Role and Function (N = 2288)

	No. of Items	Mean	S.D.
Total Score	250	185.9	17.3
Continuous Ventilation	50	37.8	4.8
Airway Care	50	37.5	3.7
Emergency Care	38	30.3	3.2
Infection Control	24	13.8	2.5
Cardiopulmonary Pharmacology	38	28.1	3.7
Pulmonary Function Testing	25	20.1	3.2
Cardiorespiratory Rehabilitation	25	18.3	2.1

The test data were further analyzed to provide performance results for the various examinee categories specified by the biographical data sheet. In Tables 4 and 5, Level II and Level III Proficiency Examination results were compared with results from the respective credentialing examinations by examinee category. A complete comparison of examinee performance by category was provided for each subtest of the examinations in an earlier report to the contracting agency.

TABLE 4

Comparison of Level II Proficiency Examination Performance with Certification Examination Performance by Examinee Category

Examinee Category	Number of Examinees	Proficiency Examination Mean	Proficiency Examination S.D.	Certification Examination Mean	Certification Examination S.D.
Total	3921	171.2	23.4	175.3	28.2
First-Timer Group	2969	172.5	24.0	176.8	29.0
Repeater Group	952	167.2	20.9	170.8	24.7
Professional Status					
Card-Pulm Technologist	428	171.0	22.5	175.3	27.0
Licensed Voc. Nurse	98	166.4	18.2	167.8	25.0
Cert. Resp. Technician	420	164.8	24.9	168.9	28.6
Licensed Pract. Nurse	215	171.0	20.8	174.5	26.0
Regis. Resp. Therapist	137	158.2	26.6	162.7	31.4
Cert. Nurse Anesth.	30	164.8	19.6	174.4	27.7
Registered Nurse	75	173.0	24.0	179.4	29.1
AMA-Approved Respirator Program					
Graduate	702	177.5	23.3	183.6	27.9
Not Graduate	2981	169.4	22.6	173.0	27.4
Now Attending	173	183.4	2[illegible]4	190.3	28.0

TABLE 4 (cont.)

Examinee Category	Number of Examinees	Proficiency Examination Mean	Proficiency Examination S.D.	Certification Examination Mean	Certification Examination S.D.
Years of Respiratory Therapy Education					
None	1130	172.0	22.6	175.7	27.5
Less than 1 year	585	171.6	23.6	175.5	28.3
1 year	642	175.8	22.2	181.2	25.9
2 years or more	1528	169.0	23.7	173.1	28.8
Highest Level of Education					
Less than high school	88	156.7	21.9	155.7	25.4
High school graduate	1353	164.0	22.1	166.2	26.6
Practical nursing school	199	170.5	19.2	172.5	24.8
Registered nursing school	40	175.6	22.3	181.2	26.3
Allied health school	206	174.0	20.4	178.3	24.9
Technical school	196	170.4	20.4	174.3	25.7
2 years or less/college	979	173.7	22.0	178.5	26.1
Over 2 years/college	532	180.5	23.7	187.9	27.6
College graduate	315	181.9	26.0	189.9	30.0
Type of Facility in Which Working					
Hospital	3732	171.0	23.3	175.1	28.0
Clinic	21	166.1	32.0	168.2	37.9
Service Company	68	178.6	19.9	184.0	23.6
Other	29	173.6	25.0	177.2	32.1
Not employed	60	176.4	25.3	181.8	29.8
Years of Respiratory Therapy Experience					
Less than 1 year	48	171.4	24.1	180.2	28.7
1 year	149	173.0	28.3	179.6	32.3
2 years	883	177.6	23.1	182.6	28.2
3 years	1142	171.3	23.6	176.0	28.5
4 years or more	1686	167.8	22.0	170.8	26.5
Experience in Intensive Care					
None	271	162.1	24.6	165.7	30.0
Under 3 months	207	167.1	25.0	173.0	29.3
3-6 months	214	171.3	23.3	174.6	28.7
7-12 months	294	173.3	24.1	178.6	29.2
1-2 years	877	175.6	22.5	180.0	27.2
2 years or more	2043	170.8	22.8	174.7	27.4
Experience in Pulmonary Labs					
None	1663	168.6	23.2	172.0	28.2
Under 3 months	686	176.0	22.3	181.1	27.6
3-6 months	352	172.7	22.7	177.2	26.5
7-12 months	244	174.6	22.6	179.5	27.5
1-2 years	479	172.4	23.1	177.2	27.9
2 years or more	483	170.0	24.5	174.4	28.1

TABLE 4 (cont.)

Examinee Category	Number of Examinees	Proficiency Examination Mean	Proficiency Examination S.D.	Certification Examination Mean	Certification Examination S.D.
Draw Blood Gases in Job					
Yes	1543	174.2	23.3	178.9	28.0
No	2319	169.2	23.1	173.0	27.9
Experience in Pediatric Therapy					
None	1540	172.1	22.9	176.4	27.8
Under 3 months	585	175.4	22.9	181.2	27.7
3-6 months	335	177.4	23.0	182.5	27.1
7-12 months	276	173.5	21.3	178.5	25.7
1-2 years	463	169.2	22.8	172.8	27.1
2 years or more	699	163.6	23.6	165.7	28.6
Experience in Card-Pulm Rehabilitation					
None	1792	171.7	23.5	175.6	28.6
Under 3 months	418	174.8	25.3	180.7	30.1
3-6 months	247	175.2	22.0	179.7	26.0
7-12 months	215	173.8	20.3	178.7	24.4
1-2 years	466	170.7	24.3	175.0	28.6
2 years or more	750	166.9	21.2	170.0	25.7
Current Job					
Part-time Instructor	23	181.5	25.8	191.6	26.5
Full-time Instructor	34	177.1	37.4	148.4	43.6
Supervisor	337	175.6	21.9	180.8	26.9
Dept. Head/Chief Therap.	204	177.9	21.7	182.4	26.7
Staff Resp. Therapist	561	170.7	25.2	175.3	30.0
Staff Resp. Technician	2367	170.7	22.5	174.5	27.2
Card-Pulm Technologist	103	173.2	23.1	177.4	27.6
Other	219	167.0	22.8	171.5	28.5
Not employed	59	173.6	25.6	178.0	30.1
Work Under Physician					
Yes	3509	171.9	23.3	176.1	28.0
No	340	164.8	21.9	168.0	26.9

TABLE 5

Comparison of Level III Proficiency Examination Performance with Registry Examination Performance by Examinee Category

Examinee Category	Number of Examinees	Proficiency Examination Mean	Proficiency Examination S.D.	Registry Examination Mean	Registry Examination S.D.
Total	2288	185.9	17.3	128.4	21.8
Professional Status					
Card-Pulm Technologist	123	184.8	16.3	128.9	19.4
Licensed Voc. Nurse	23	183.8	13.2	124.5	14.8

TABLE 5 (cont.)

Examinee Category	Number of Examinees	Proficiency Examination Mean	Proficiency Examination S.D.	Registry Examination Mean	Registry Examination S.D.
Professional Status (cont.)					
Cert. Resp. Technician	1094	186.6	16.4	128.2	21.5
Licensed Pract. Nurse	26	174.3	14.8	116.2	13.1
Regis. Resp. Therapist	199	193.7	19.0	137.6	25.2
Cert. Nurse Anesthetist	14	185.3	15.8	127.6	21.5
Registered Nurse	37	185.5	16.8	124.9	19.5
AMA-Approved Respiratory Program					
Graduate	2212	185.7	17.2	128.2	21.6
Not Graduate	69	190.4	19.6	133.3	25.6
Now Attending	5	198.0	13.7	150.6	21.6
Years of Respiratory Therapy Education					
None	28	190.1	21.6	136.0	24.4
Less than 1 year	160	189.5	17.8	136.6	21.7
1 year	161	187.4	19.9	133.2	22.4
2 years or more	1938	185.4	16.9	127.3	21.5
Highest Level of Education					
Less than high school	0	0.0	0.0	0.0	0.0
High school graduate	12	177.0	29.6	122.0	28.6
Practical nursing school	3	170.3	14.4	116.3	10.3
Registered nursing school	15	181.5	19.1	126.4	19.8
Allied health school	85	183.5	14.1	124.5	18.7
Technical school	51	179.8	15.6	119.6	17.1
2 years or less/college	367	182.3	16.5	122.0	21.1
Over 2 years/college	1035	186.3	17.0	128.9	21.3
College graduate	720	188.1	17.7	132.4	22.5
Type of Facility in Which Working					
Hospital	2154	185.4	17.2	128.0	21.5
Clinic	12	194.1	11.3	138.6	18.3
Service company	24	186.2	11.9	120.4	21.9
Other	76	198.5	17.2	144.2	24.4
Not Employed	20	186.3	14.7	125.4	19.3
Years of Respiratory Therapy Experience					
Less than 1 year	223	183.7	16.5	127.2	20.9
1 year	336	185.0	17.3	130.1	22.0
2 years	573	185.6	16.1	128.6	20.6
3 years	478	185.3	17.5	127.9	21.0
4 years or more	678	187.6	18.3	128.3	23.5
Experience in Intensive Care					
None	21	175.0	19.9	115.6	21.9
Under 3 months	94	183.6	21.0	126.8	21.8
3-6 months	232	184.2	17.0	128.8	21.8
7-12 months	388	184.0	17.3	128.6	20.3
1-2 years	683	185.6	15.8	128.5	21.4
2 years or more	868	187.9	17.7	128.7	22.7

TABLE 5 (cont.)

Examinee Category	Number of Examinees	Proficiency Examination Mean	Proficiency Examination S.D.	Registry Examination Mean	Registry Examination S.D.
Experience in Pulmonary Labs					
None	690	184.6	17.6	127.4	21.5
Under 3 months	751	186.0	16.8	129.0	21.6
3-6 months	243	184.2	18.7	126.7	22.8
7-12 months	189	187.7	15.7	129.3	21.4
1-2 years	236	187.1	16.8	130.1	21.7
2 years or more	176	189.1	17.7	129.3	23.4
Draw Blood Gases in Job					
Yes	1305	187.0	16.5	130.2	21.5
No	965	184.4	18.1	126.2	22.2
Experience in Pediatric Therapy					
None	803	185.8	17.3	128.5	21.4
Under 3 months	625	186.5	17.2	129.8	22.0
3-6 months	288	185.1	17.6	128.5	21.8
7-12 months	211	186.3	15.5	129.2	20.6
1-2 years	205	185.9	17.3	126.0	22.7
2 years or more	152	184.1	19.3	124.7	23.8
Experience in Card-Pulm Rehabilitation					
None	1135	186.6	17.0	130.1	21.7
Under 3 months	438	187.1	16.7	129.7	21.7
3-6 months	174	184.7	19.4	126.3	23.6
7-12 months	151	185.8	15.6	125.9	19.9
1-2 years	200	183.9	18.1	126.4	22.5
2 years or more	183	181.6	18.0	121.4	20.4
Current Job					
Part-time Instructor	45	193.1	15.6	135.6	21.0
Full-time Instructor	125	199.9	13.8	147.8	19.8
Supervisor	406	187.5	16.0	129.1	21.5
Dept. Head/Chief Therap.	319	187.3	16.3	126.7	21.2
Staff Resp. Therapist	1084	183.4	17.1	126.4	21.1
Staff Resp Technician	198	181.9	19.2	125.9	22.2
Card-Pulm Technologist	19	188.8	16.5	134.1	24.7
Other	72	188.6	18.2	131.1	22.8
Not Employed	18	187.8	14.4	127.9	18.1
Work Under Physician					
Yes	2155	186.1	17.2	128.7	21.8
No	113	182.0	17.7	124.4	22.1

Statistical comparisons of test and subtest performance on the Proficiency and Credentialing Examinations are presented in Tables 6 and 7. Table 6 includes all product-moment correlations among the various scores and subscores of the Level II Proficiency Examination and the CRTT Certification Exam. Similarly, correlations among the scores and subscores of the Level III Proficiency Examination and the Written Registry Exam are provided in Table 7.

TABLE 6

Correlation of Scores Between the Level II Proficiency Examination and the Certification Examination (N = 3921)

Level II Proficiency Examination	Certification Examination					
	Total	Life Sci	Phys. Sci	Gas Adm	PPB Vent	Card-Pulm
Total	0.9013	0.7909	0.8006	0.8064	0.8240	0.7632
IBBP	0.7855	0.6837	0.7035	0.7163	0.7712	0.6910
Hem/Aer Ther	0.7216	0.6182	0.6440	0.6696	0.6803	0.5926
Gas Ther	0.8282	0.7658	0.7357	0.7763	0.7489	0.6926
Pulm Drn	0.7076	0.6294	0.6271	0.6204	0.6561	0.6506
C-R Resusc	0.5731	0.4773	0.5015	0.4996	0.5475	0.5429
C-R Drug Adm	0.7420	0.6545	0.6982	0.6752	0.6775	0.6590
Infect Contr	0.4563	0.4000	0.4114	0.4058	0.4186	0.4147

Summaries of available item analysis statistics for the two Proficiency Examinations are presented in Tables 8 and 9. The tables specify the numbers of items with p-values in each portion of the difficulty range (0-1.00). The tables also provide the distribution of point-biserial correlation coefficients between item performance and total score on the respective Proficiency Examinations. In addition, correlations between item performance and performance on the Certification Examinations were available for the Level II test and are included in Table 8. The contracting agency was provided a complete listing of item statistics for the two Proficiency Examinations in an earlier report.

TABLE 8

Summary of Item Analysis Statistics for Level II Proficiency Examination

Range	Difficulty	Item-Prof. Test Correlation	Item-Cert. Test Correlation
0.91-1.00	45	0	0
0.81-0.90	46	0	0
0.71-0.80	36	0	0
0.61-0.70	38	0	0
0.51-0.60	33	0	0
0.41-0.50	21	8	8
0.31-0.40	16	61	44
0.21-0.30	12	83	81
0.11-0.20	3	69	79
0.01-0.10	0	25	30
≤ -0.00	0	4	8

TABLE 7

Correlation of Scores Between the Level III Proficiency Examination and the Registry Written Examination (N = 2288)

Level III Proficiency Examination	Written Registry Examination									
	Total	Clinical	Physio	Pulm Fnc	Therapy	Anatomy	Chemistry	Pharma	Bacter	Equpt
Total	0.8141	0.6854	0.7007	0.6485	0.6907	0.4899	0.4971	0.5603	0.4680	0.6459
Cont Vent	0.7384	0.6057	0.6459	0.5831	0.6218	0.4418	0.4498	0.4715	0.3892	0.5618
Air Care	0.6033	0.5140	0.4960	0.4544	0.5100	0.3541	0.3366	0.4250	0.3345	0.4810
Emergency	0.6040	0.5017	0.4997	0.4793	0.5030	0.3681	0.3591	0.3951	0.3731	0.4896
Infection	0.4660	0.3901	0.3939	0.3406	0.4046	0.2656	0.2675	0.3229	0.3173	0.3755
C-R Pharm	0.6969	0.5803	0.5877	0.5244	0.5759	0.4031	0.4358	0.5235	0.4082	0.5465
Pulm Test	0.6783	0.5360	0.5871	0.6091	0.5425	0.4287	0.4322	0.4141	0.3696	0.5196
C-R Rehab	0.3197	0.2884	0.2614	0.2223	0.2808	0.1619	0.1814	0.2488	0.2279	0.2239
Sit Sets	0.6901	0.5077	0.5301	0.4767	0.5010	0.3549	0.3637	0.4235	0.3499	0.4895

TABLE 9

Summary of Item Analysis Statistics
for Level III Proficiency Examination

Range	Difficulty	Item-Prof. Test Correlation
0.91-1.00	74	0
0.81-0.90	60	0
0.71-0.80	31	0
0.61-0.70	21	0
0.51-0.60	24	0
0.41-0.50	11	1
0.31-0.40	10	33
0.21-0.30	11	75
0.11-0.20	7	99
0.01-0.10	1	34
≤-0.00	0	8

The complete frequency distributions of scores obtained on the Proficiency Examinations and on the Certification Examination were also contained in the earlier report. For the present report, these data were summarized by combining the scores within each ten-point interval and are presented in Table 10 as frequencies and proportions of examinee scores falling in each interval. Written Registry Examination scores were not provided to the Association by ETS and could not be summarized for inclusion in the table.

TABLE 10

Frequency Distribution of Examination Scores

	Level II				Level III	
	Prof. Exam		Cert. Exam		Prof. Exam	
Interval	No.	Prop.	No.	Prop.	No.	Prop.
0-10	0	.0000	0	.0000	0	.0000
11-20	1	.0003	0	.0000	0	.0000
21-30	0	.0000	1	.0003	0	.0000
31-40	0	.0000	0	.0000	0	.0000
41-50	0	.0000	0	.0000	0	.0000
51-60	0	.0000	0	.0000	0	.0000
61-70	0	.0000	2	.0005	0	.0000
71-80	5	.0013	2	.0005	1	.0004
81-90	3	.0008	6	.0015	0	.0000
91-100	12	.0031	16	.0041	3	.0013
101-110	21	.0054	48	.0122	1	.0004
111-120	59	.0150	67	.0171	1	.0004
121-130	113	.0288	149	.0380	6	.0026
131-140	183	.0467	195	.0497	16	.0070
141-150	284	.0724	228	.0581	34	.0149
151-160	463	.1181	351	.0895	107	.0468
161-170	658	.1678	491	.1252	233	.1018
171-180	734	.1872	575	.1466	395	.1726
181-190	608	.1551	628	.1602	524	.2290
191-200	403	.1028	428	.1092	498	.2177
201-210	212	.0541	332	.0847	347	.1517
211-220	125	.0319	213	.0543	110	.0481
221-230	36	.0092	148	.0377	11	.0048
231-240	1	.0003	38	.0097	1	.0004
241-250	0	.0000	3	.0008	0	.0000

To facilitate comparison of score distributions among the examinations, the data of Table 10 are presented graphically in Figures 1 and 2. Although individual scores obtained on the Written Registry Examination were not available from ETS, the score distribution was approximated from the group mean and standard deviation and is included in Figure 2.

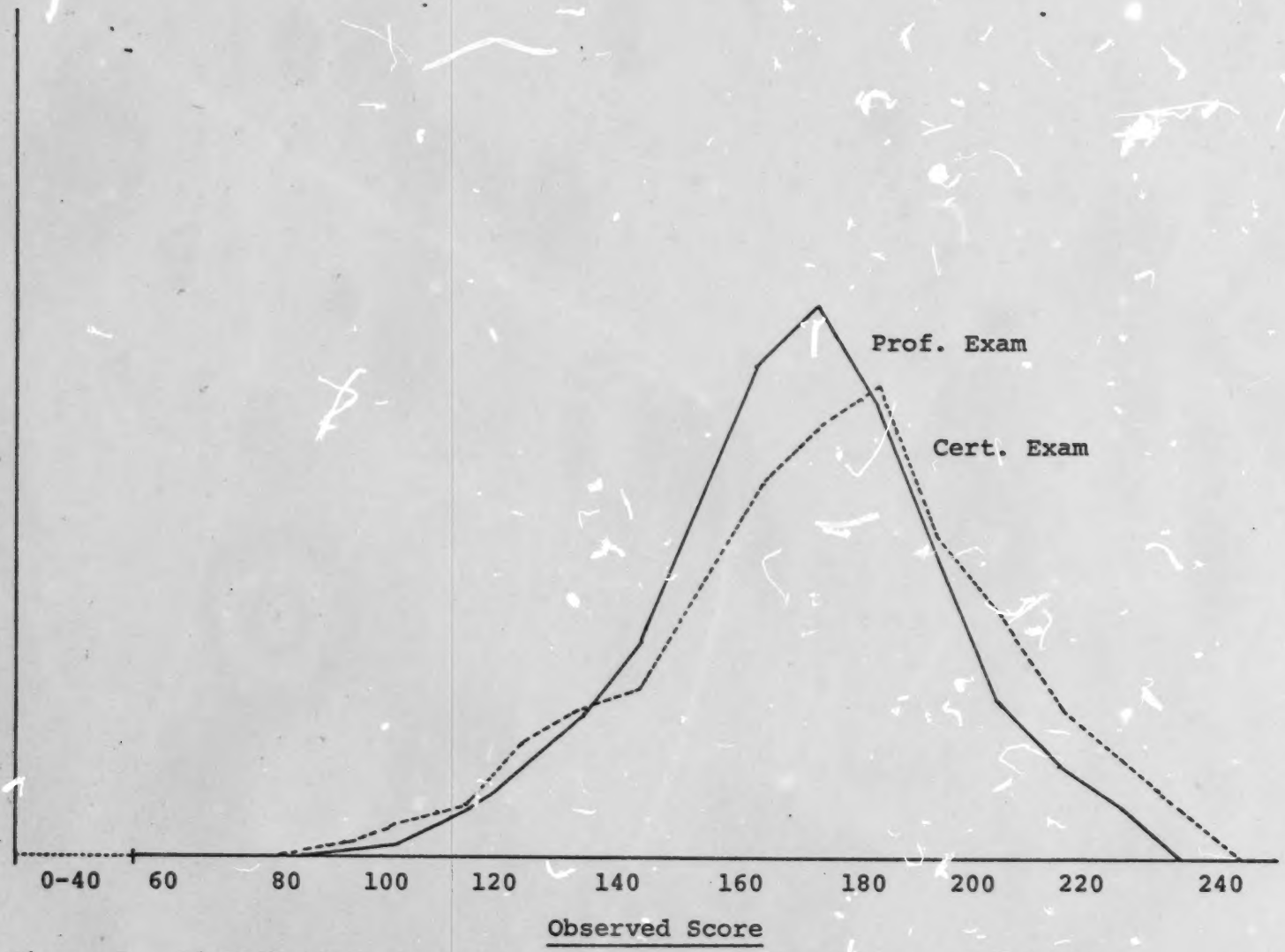

Figure 1. Distribution of scores on the Level II Proficiency Examination and the CRTT Certification Examination.

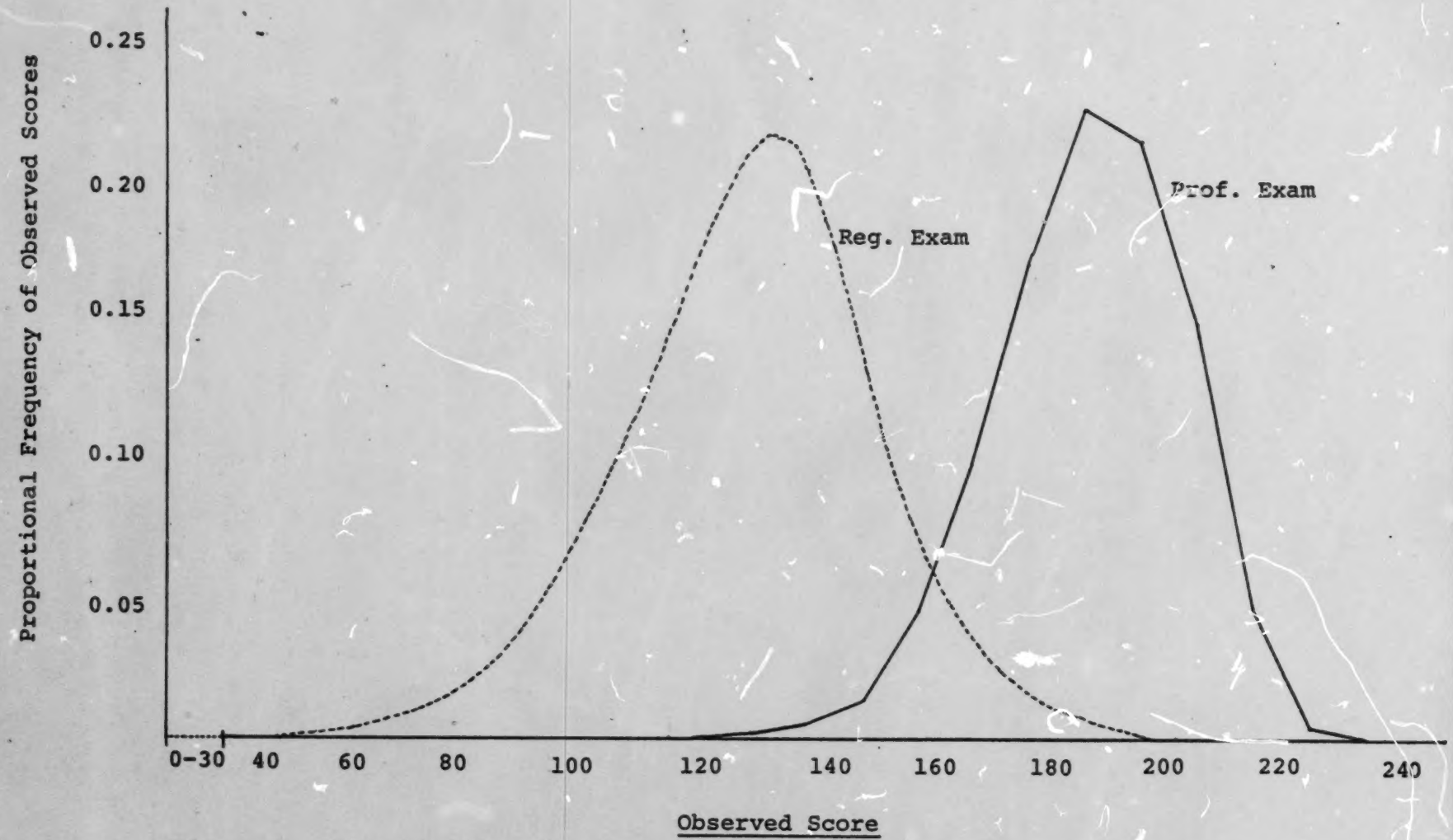

Figure 2. Distribution of scores on the Level III Proficiency Examination and estimated distribution of scores on the ARRT Registry Examination.

DISCUSSION

It is important that discussion of the results of the validation study begin with a clarification of the Association's past and current understanding of the process of proficiency testing. From the beginning, it was assumed that proficiency is most appropriately measured by criterion-referenced procedures. Due to the embryonic status of criterion-referenced measurement (CRM), however, the Association has been highly dependent upon external advice for its understanding of CRM methodology. Unfortunately, this dependency has resulted in a number of activities which in retrospect appear to have been ill-advised. The Association willingly accepts its share of the responsibility for errors of the past and welcomes the opportunity to initiate activities which will improve the state of the art of proficiency testing.

Since the introduction of the term "criterion-referenced" by Glaser (1963), a wide variety of definitions and interpretations of the term has appeared in the literature. There was a time when one of the few areas of agreement was that CRM differs from norm-referenced measurement and that interpretation of an individual's criterion-referenced test (CRT) score should be independent of reference to the performance of others. It is now generally agreed that a CRT score must be directly interpretable in terms of some prespecified capability or class of behavior. It is this clearly defined capability, usually operationalized by a behavioral objective, that is "referenced" by a criterion-referenced test.

During Phase I of the project, the Association attempted to define the entry level capabilities of respiratory therapy personnel. This delineation of roles and functions represented a substantial improvement over existing statements of personnel competence. In retrospect, however, these statements did not provide sufficient behavioral specificity for the derivation of criterion-referenced proficiency examinations. This criterion problem was magnified by procedures subsequently employed in the development and refinement of the proficiency examinations. Items were generated for the exams by methods which are more appropriately utilized in the construction of norm-referenced tests. Although CRM methodology remains in the early stages of development, many of these problems could have been avoided by using procedures which had appeared in the literature prior to the time of test development. Hively, Patterson, and Page (1968) described procedures for sampling items from a population specified in advance by item forms. Kriewall (1969) presented a thorough development of the theory and methodology related to the use of systematic item-sampling procedures for CRT construction. The use of amplified objectives by Popham (1972) and of domain-referenced test construction procedures by Millman (1973) provided additional advances in CRM methodology which should have been considered during the test construction phase of the project.

In addition to using norm-referenced procedures for constructing the Proficiency Examinations, field testing and evaluation of the exams during Phase II also followed approaches generally employed with norm-referenced tests. Since these same procedures were recommended to the Association for implementation in the present study, specific limitations of the methodology will be presented in the discussion that follows. In both instances, however, the use of procedures which permit criterion-referenced interpretations (Fremer 1972) should have received consideration as alternatives.

Table 1 presented summary statistics for the Proficiency and Certification Examinations. The concepts of mean, variance, standard deviation, and standard error have limited meaning outside of a norm-referenced context. These concepts describe relationships among sets of scores. It should be noted, however, that the statistical values for the Level II Proficiency Examination closely approximate those of the norm-referenced competency examination. This finding strongly suggests that the Proficiency Exam was more norm-referenced in character than criterion-referenced.

Although based on variance among a set of scores, an internal consistency estimate of reliability, such as KR20 or KR21, may be appropriate for criterion-referenced tests under certain conditions. If all items on an examination are measuring the same defined capability, high inter-item correlation would be expected. In fact, Graham (1974) consistently obtained KR20 reliabilities above 0.9 for randomly generated ten-item subtests. In the present situation, however, the reliability estimates were provided for the total 250-item test rather than for the individual subtests. Since the collection of subtests was supposed to be measuring several different capabilities, homogeneity among subtests would not be expected and the use of an internal consistency index for the total test was probably inappropriate. Separate KR20 values for subtests of a CRT would be more readily interpretable.

Means and standard deviations for the subscales of the Proficiency Examinations were presented in Tables 2 and 3. As indicated above, these statistics have little meaning for criterion-referenced instruments. If each subtest measured a specific, defined competency and if each subtest had a pre-established passing score, the proportion of examinees demonstrating minimum proficiency would provide a more meaningful index. Regrettably, the tests evaluated in the present study did not satisfy either of these two conditions, i.e., subtests were not based on adequately defined competencies and a priori passing scores were not established.

The previous comments regarding Tables 2 and 3 are equally applicable to Tables 4 and 5. Although it would be inappropriate to compare Proficiency Examination means and standard deviations among various examinee categories, such a comparison would be defensible for the two credentialing exams. However, since no

particular differences were hypothesized in advance and since the probability of Type I errors in a post hoc analysis involving this many variables would be extremely high, no statistical comparisons were performed.

The Proficiency Exam subscores (Tables 4 and 5) appear to have appreciably lower standard deviations than those of the credentialing exams. Although the standard deviations for CRTs are sometimes expected to be lower than for NRTs, an alternative explanation for the differences observed in the present study should be considered. Norm-referenced instruments such as the Certification and Registry Exams are typically subjected to an iterative revision process which tends to insure greater variance in test scores by revising or discarding items that do not adequately discriminate among examinees. Thus, the observed differences in standard deviations could be explained by differences in the degree of refinement of the respective tests. Although the Level III Proficiency Exam also differed substantially from the Registry Exam in difficulty, the overall performance of the two Proficiency Exams appeared quite similar to their norm-referenced analogs.

The product moment correlation coefficients presented in Tables 6 and 7 indicate a high probability of a relationship existing between each of the Proficiency Examinations and the corresponding credentialing exam. The probability that: (a) a correlation of 0.9013 between the Level II Proficiency Exam and the Certification Exam; and (b) that a correlation of 0.8141 between the Level III Exam and the Registry Exam are chance occurences is exceedingly small. In fact, of the 138 correlation coefficients in the two tables, the smallest coefficient of 0.1619 between Cardiorespiratory Rehabilitation (Level III Proficiency Exam) and Anatomy (Registry Exam) is significantly different than zero at the .001 level. The fact that the Proficiency Examinations can account for such a high proportion of the variance on the credentialing exams suggests that the two sets of examinations are probably measuring the same variables in much the same way. If this is correct, the Proficiency Examinations could probably serve as alternate forms of the corresponding credentialing exams. Moreover, the correlations between the total tests (0.9013 and 0.8141) exceed the alternate forms reliability coefficients of many tests that are constructed to be parallel.

Tables 8 and 9 summarize the item difficulty values for the two Proficiency Examinations. Item difficulty is a meaningful index for criterion-referenced tests. The difficulty of an item that is truly referenced to an objective, however, is sometimes considered to be a function of the learning state of the examinee rather than a function of item characteristics. Under these circumstances, the p-value (difficulty) of an item depends upon the relative number of examinees in the learned and unlearned states relative to the objective being measured. With this perspective, all items measuring the same competency should display relatively similar or homogeneous difficulty values. Since the items of the

Proficiency Examinations were not identified by subtest, a direct determination of the existence of homogeneity of item difficulties was not possible. Such homogeneity appears unlikely, however, because: (a) the items were not derived from objectives, (b) the subtests were designed to measure broad, general competencies with little restriction upon permissible item characteristics, and (c) subtest characteristics were quite similar while item difficulties for the collection of subtests were dissimilar.

Tables 8 and 9 also summarize the point biserial correlation coefficients between items and total test performance. As in the case of item difficulties, these statistics are most meaningful when they are calculated within subtests which measure a given competency rather than for the test as a whole. It should be noted that, for Level II, item performance was correlated with Certification Exam scores as well as Proficiency Exam scores (Table 8). The similarity of these two sets of coefficients further substantiates the previous assertion that the two instruments are probably measuring the same variables.

Additional similarities and certain differences between the Proficiency and Credentialing Examinations are apparent from examining Table 10 and Figures 1 and 2. Scores from each of the instruments displayed characteristics of a normal distribution. Although slightly more leptokurtic, the score distribution of the Level II Proficiency Exam was nearly identical to the Certification Exam in other respects. The distribution of scores for the Level III Proficiency Exam was also more leptokurtic than its credentialing analog. In addition, the Proficiency Exam showed considerable negative skewness compared to the Registry Exam.

Leptokurtosis and skewness are acceptable but unessential characteristics of CRT score distributions. Kurtosis is dependent upon the homogeneity of both items and examinees. Skewness is a function of the mean item difficulty and depends on the nature of the items and the capabilities of the examinees. In the present study, the score distributions for the Proficiency Exams were not judged to differ substantially from those of the corresponding credentialing exams. Although less than perfect, the Proficiency Exam distributions appeared to be typical examples of the score distributions displayed by many norm-referenced examinations.

To summarize the discussion, the Association believes:

1. The delineation of roles and functions did not provide an adequately defined basis for the generation of criterion-referenced Proficiency Examinations.
2. The procedures employed in the construction of the Level II and Level III Proficiency Examinations were more appropriate for a norm-referenced test and did not produce instruments that differed substantially from existing credentialing examinations.
3. Even if the Proficiency Examinations could be considered criterion-referenced in character, procedures employed

in the analysis of the examination results were generally inappropriate for such tests.

4. More acceptable procedures for analyzing criterion-referenced test data are available. These procedures could be employed in the present situation, but in light of such factors as: (a) the procedures by which the tests were constructed; (b) the results of the present analysis; and (c) the cost associated with further analysis, subjecting these data to further analytical investigation does not appear warranted.

The Association is presently in the process of identifying and defining with greater specificity the entry level competencies of the profession. In addition, the Association is extensively upgrading its understanding and proficiency in the area of criterion-referenced measurement. As evidence of this endeavor, a review of criterion-referenced measurement methodology is attached to this report as a supplement (Appendix D). It is believed that these efforts will ultimately result in a product which can demonstrate its quality as a criterion-referenced proficiency test and which will serve as a model for other allied health professions.

RECOMMENDATIONS

One long-range goal of the federally-funded educational research underway by the Association and represented in part by this project, is the establishment of a credentialing examination system that will better serve the profession and the public by accurately discriminating between those who are actually competent and those who are not within the several competency areas in respiratory therapy. As criterion-referenced measurement currently holds the highest promise for evolving such a system, the two-phase project producing the Levels I and II Proficiency Examinations specified criterion-referenced exams as the desirable products.

Regretfully, intervening experience has demonstrated that this expectation was not substantially fulfilled. This fact leaves both the Association and the Division of Associated Health Professions in an awkward position as regards these examinations. It may be possible, by methods not yet proposed, to convert the Proficiency Examinations into criterion-referenced instruments, but not likely.

The most efficient course leading to CR Proficiency Exams in respiratory therapy would probably be to start again. In fact, such work is now underway in the AART/HRA project 231-75-0213, where a "definition" of respiratory therapy competence is in development to supplant the DELINEATION OF ROLES document. Further, based on selected competency "domains" within this definition, criterion-referenced test specifications will be written under the direction of W. James Popham. These specifications will then be used as a basis for developing at least one evaluative simulation for pilot-testing by the NBRT as a possible alternative assessment form within the credentialing system.

Therefore, based on the considerations treated in this report, the following are recommended:

1. That the Proficiency Examinations developed under NO1-AH-34062 not be subjected to further review or revision at federal expense;
2. That these examinations, including the entire item pool and any other materials, produced under NO1-AH-34062 now being held by Psychological Corporation, be turned over to the National Board for Respiratory Therapy for use at their discretion;
3. That the Division of Associated Health Professions submit a Request for Proposal to the National Board for Respiratory Therapy as the sole source agency to continue the research underway in HRA 231-75-0213 in criterion-referenced credentialing examination development.

REFERENCES

Fremer, J. Criterion-referenced interpretations of survey achievement tests. Test Development Memorandum 72-1, January, 1972, Educational Testing Service, Princeton, New Jersey

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Graham, D.L. An empirical investigation of the application of criterion-referenced measurement to survey achievement testing. Unpublished doctoral dissertation, Florida State University, 1974.

Hively, W., Patterson, H.L., & Page, S.H. A "universal-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Kriewall, T.E. Application of information theory and acceptance sampling principles to the management of mathematics instruction. Technical Report No. 103, October, 1969, Wisconsin Research and Development Center, Madison, Wisconsin.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-215.

Popham, W.J. Selecting objectives and generating test items for objectives-based tests. Los Angeles: Instructional Objectives Exchange, 1972.

THE NATIONAL BOARD FOR RESPIRATORY THERAPY, INC.
(formerly American Registry Of Inhalation Therapists)

Executive Office
Suite 124
1900 W 47th St.
Westwood, Kansas 66205
[illegible]

EXECUTIVE COMMITTEE
President
Robert H. Miller, (AART)

Vice President
Hurley L. Motley, MD, (ACCP)

Treasurer
Allan B. Saposnick, (AART)

Secretary
LeRoy Misuraca, MD, (ASA)

Past President
Duncan A. Holaday, MD, (ASA)

Walter J. O'Donohue, Jr., MD, (ATS)

Jimmy A. Young, (AART)

MEMBERS OF THE BOARD
John B. Berte, MD, (ACCP)

William W. Burgin, Jr., MD, (ATS)

Frederick W. Cheney, MD, (ASA)

Herbert T. Constantine, MD, (ATS)

Frank Dick, (AART)

Sister Bernice Ebner, (AART)

William Gardiner, (AART)

Robert Glass, (AART)

Gary Lynch, (AART)

Jerome J. Maurizi, MD, (ACCP)

Edward H. Morgan, MD, (ACCP)

John T. Sharp, MD, (ATS)

Louis M. Sinopoli, (AART)

Carole Trout, (AART)

George West, (AART)

Roy D. Wilson, MD, (ASA)

Harrel D. Ziecheck, (AART)

STAFF
Clifford D. Bryan
Executive Director

Steven K. Bryant
Executive Secretary

Robert M. Lawrence, MD
Director–Oral Examinations

Dear Certification Examination Candidate:

There will be a change in the format of the CRTT Written Examination on May 10, 1975. The examination will consist of two parts, and will require your attendance from approximately 8:00 A.M. to 4:00 P.M. on your testing day (one hour break for lunch).

The first part of the examination in the morning will consist of approximately 200 questions, objective and multiple-choice type. The second part is structurally different, stressing behavioral goals.

We ask that you take this two-part examination to assist us in the growth and development of the credentialing process. We suggest that you endeavor to answer all of the questions to the best of your ability.

We hope that asking you to participate in this unique opportunity will not create any hardship or additional stress for you to bear. This is not our intention, and we sincerely request your cooperation and understanding.

On behalf of the Technician Written Examination Committee, we wish you success.

Sincerely,

Sister Bernice Ebner

Sister Bernice Ebner, ARRT
Chairman, Technician
Written Examination Committee

sbe/dr
2/75

Sponsored By: American Society of Anesthesiologists
American College of Chest Physicians
American Association for Respiratory Therapy
American Thoracic Society

Dear Written Exam Candidate:

There will be a change in the format of the ARRT Written Exam in March 1975. The exam will consist of two parts, and will require your attendance from approximately 8:00 A.M. to 4:00 P.M. on your testing day (one hour break for lunch).

The first part of the exam in the morning will consist of 200 test items and will be the major basis for your grade. The second part of the exam is structurally different, stressing behavioral goals, and can only raise your grade on Part 1 of the exam.

We ask that you take this two-part exam in order to assist us in the growth and development of the credentialing process. Your performance on Part 2 of the exam will in no way lower your grade. We suggest, however, that you endeavor to answer all of the questions to the best of your ability, since your score on Part 2 could raise your grade on Part 1.

We hope that asking you to participate in this unique opportunity will not create any hardship or additional stress for you to bear. This is not our intention, and we sincerely request your cooperation and understanding.

On behalf of the Written Exam Committee, we wish you success.

Very truly yours,

Louis M. Sinopoli

Louis M. Sinopoli, ARRT, Chairperson
ARRT Written Exam Committee
of the NBRT

CANDIDATE QUESTIONNAIRE

NAME (Print) _____ SOCIAL SECURITY NO. _____
(Last) (First) (Middle)

PERMANENT ADDRESS _____ YEAR OF BIRTH _____
(Street) (City) (State) (Zip)

DIRECTIONS: Please answer all of the following questions as accurately as possible. Indicate the number of your answer in the appropriate space to the left of each question. There should be one number recorded in each space.

What professional status applies to you? (Indicate as follows: 1 = Yes; 2 = No)

1. _____ Cardio-Pulmonary Technologist (CPT)
2. _____ Licensed Vocational Nurse (LVN)
3. _____ Certified Respiratory Therapy Technician (CRTT)
4. _____ Licensed Practical Nurse (LPN)
5. _____ American Registered Respiratory Therapist (ARRT)
6. _____ Certified Registered Nurse Anesthetist (CRNA)
7. _____ Registered Nurse (RN)
8. _____ Are you a graduate of an AMA-approved respiratory therapy program?
   1. Yes
   2. No
   3. Am presently attending
9. _____ How many years of respiratory therapy education have you had?
   1. None
   2. Less than 1 year
   3. 1 year
   4. 2 years or more
10. _____ What is the most advanced level of education you have completed?
    1. Less than high school
    2. High school graduate
    3. Practical nursing school
    4. Registered nursing school
    5. Other allied health school
    6. Technical school
    7. 2 years or less of college
    8. More than 2 years of college
    9. College graduate
11. _____ In what type of faculty are you presently employed?
    1. Hospital
    2. Clinic
    3. Service company
    4. Other
    5. Not employed
12. _____ How many years of respiratory therapy experience have you had?
    1. Less than 1 year
    2. 1 year
    3. 2 years
    4. 3 years
    5. 4 years or more
13. _____ How much work experience have you had in Intensive Care Units?
    1. None
    2. Less than 3 months
    3. 3-6 months
    4. 7-12 months
    5. 1-2 years
    6. 2 or more years
14. _____ How much work experience have you had in pulmonary function laboratories?
    1. None
    2. Less than 3 months
    3. 3-6 months
    4. 7-12 months
    5. 1-2 years
    6. 2 years or more
15. _____ Do you draw blood gases as part of your present job?
    1. Yes
    2. No
    3. Not employed
16. _____ How much work experience have you had exclusively in pediatric respiratory therapy?
    1. None
    2. Less than 3 months
    3. 3-6 months
    4. 7-12 months
    5. 1-2 years
    6. 2 years or more
17. _____ How much work experience have you had exclusively in cardio-pulmonary rehabilitation?
    1. None
    2. Less than 3 months
    3. 3-6 months
    4. 7-12 months
    5. 1-2 years
    6. 2 years or more
18. _____ What is your current job?
    1. Part-time instructor in respiratory therapy
    2. Full-time instructor in respiratory therapy
    3. Supervisor
    4. Department head/chief therapist
    5. Staff respiratory therapist
    6. Staff respiratory technician
    7. Cardio-pulmonary technologist
    8. Other
    9. Not employed
19. _____ Are you currently working under the direction of a physician?
    1. Yes
    2. No
    3. Not employed

THE PSYCHOLOGICAL CORPORATION
INCORPORATED IN 1921
757 THIRD AVENUE
NEW YORK, N.Y. 10017
(212) 754-3500

January 6, 1975

Mr. William Johnson
American Association for Respiratory Therapy
7411 Hines Place
Dallas, Texas 75235

Dear Bill:

This letter is to summarize the services which the Professional Examinations Division of The Psychological Corporation will provide to the American Association for Respiratory Therapy in conjunction with the administration of the Level II and Level III proficiency examinations for respiratory therapy personnel.

Certification/Level II Examination

1. The Professional Examinations Division will edit, print, and administer to Certification candidates a questionnaire on their education, experience, and other relevant background. This questionnaire, taking approximately 15 minutes to administer, is to be developed by AART and delivered to The Psychological Corporation no later than March 1, 1975, if it is to be included.

2. The Professional Examinations Division will arrange for administration of the Level II examination as Part II of the Certification Examination scheduled for May 10, 1975, and reimburse Examiners for their services.

3. The Professional Examinations Division will notify candidates of the time and place of testing.

4. The Professional Examinations Division will print and ship required materials (questionnaires, test booklets, answer sheets, directions for administration, etc.) to its examiners and will be responsible for the return and checking in of test materials, scoring the answer sheets, compiling responses to the questionnaires, and preparing the following data analyses:

a. Means and standard deviations on Level II Total Test and areas in terms of questionnaire responses.

b. Correlations among scores on the Level II and Certification Examination based on (1) first-time candidates and (2) rewriting candidates.

c. Item analysis of the Level II examination yielding difficulty and discrimination indices and showing the point biserial correlations between item performance and performance on (1) the Certification Examination and (2) Level II examination.

Registration/Level III Examination

1. The Professional Examinations Division will edit and print 3500 copies of a questionnaire on the education, experience, and relevant background of Registration candidates. This questionnaire, taking approximately 15 minutes to administer, is to be developed by AART and delivered to The Psychological Corporation by January 13, 1975, if it is to be included.

2. The Professional Examinations Division will print and ship the following materials to the Educational Testing Service (ETS):

   a. 3500 Level III test booklets

   b. 3500 Level III answer sheets

   c. 3500 questionnaires

   d. 300 directions for administration

3. Assuming materials (test booklets, answer sheets, questionnaires, directions for administration) are returned to The Psychological Corporation within four weeks following the March 1, 1975, test date, the Professional Examinations Division will score the Level III answer sheets, compile responses to the questionnaires, and prepare the following:

a. Means and standard deviations on Level III Total Test and areas in terms of questionnaire responses.

b. Item analysis on the Level III examination yielding difficulty and discrimination indices and showing the point biserial correlations between item performance and performance on the Level III examination.

(NOTE: It is our understanding that the National Board for Respiratory Therapy (NBRT) will not release the written scores from the Registration Examination for purposes of this project. Therefore, the correlation and item analyses relating the Level III and Registration Examinations cannot be performed as part of The Psychological Corporation's services. If written permission is received from Dr. Robert Conant or his representative, Level III Total and area scores could be provided to ETS.)

Cost

1. Based on approximately 5,000 candidates taking the Certification/Level II examinations, the cost for the described services would be $3.00 per candidate tested plus $30.00 per testing center.

2. Based on supplying materials for 3500 candidates taking the Registration/Level III examination, the cost for the described services will be $3.00 per candidate tested plus $1.00 for each of the 3500 candidates not tested for whom materials are supplied.

Invoices for the described services will be rendered directly to the American Association for Respiratory Therapy. The Psychological Corporation will have no financial responsibility for those services rendered by the Educational Testing Service in connection with this project.

Mr. William Johnson - 4. January 6, 1975

In order that we may have a record of your agreement with the services outlined, please sign the enclosed copy of this letter and return it to us no later than January 15, 1975. If written response is not received by this date, it may be impossible to meet the testing requirements for the Registration/Level III phase of the project.

Cordially,

Sallyann Henry, Ph.D.
Assistant Director
Professional Examinations Division

Agreed to: _____
Signature

American Association for Respiratory Therapy

Date: _____

SH:dk
encl
cc: Sister Bernice Ebner
Dr. Robert Conant

EDUCATIONAL TESTING SERVICE PRINCETON, N.J. 08540

Area Code 609
921 - 9000
CABLE-EDUCTESTSVC

Russell W. Martin, Jr.
Assistant Treasurer

January 29, 1975

Mr. William W. Johnson
Project Director
American Association for
Respiratory Therapy
7411 Hines Place
Dallas, Texas 75235

Dear Mr. Johnson:

This letter will confirm contractual arrangements between the American Association for Respiratory Therapy (AART) and Educational Testing Service (ETS) for ETS services in connection with the pilot administration of the Respiratory Therapy Proficiency Test. These services are to be performed between February 1, 1975 and April 15, 1975.

The multiple-choice examination will be administered by ETS in the afternoon of March 1, 1975 to candidates who will have taken the National Board of Respiratory Therapy's Therapist Level Written Examination in the morning administration. These candidates will be asked to complete a biographical questionnaire. The examination booklets used for this afternoon administration are to be entitled "National Board for Respiratory Therapy Written Examination - Part II."

The responsibilities of AART will be to instruct the Psychological Corporation to:

1. deliver to ETS, no later than February 1, 1975:

   a. 2,600 copies of the Proficiency Test booklets and answer sheets;
   b. 2,600 copies of the biographical questionnaire; and
   c. 180 copies of a manual of directions for administering the Proficiency Test;

2. prepare for the receipt and processing of the used answer sheets and biographical questionnaires, to be shipped by ETS as provided below; and

3. deliver to ETS, not later than two weeks after receipt of the last shipment of used answer/sheets and questionnaires from ETS, a punched and interpreted tab card containing the Proficiency Test total score and subscores, for each candidate, identified by name only.

Specifically, ETS will:

1. upon receipt of the testing materials from the Psychological Corporation, store these materials in a secure location until they are repacked for shipment to the test centers;

2. repack and ship to approximately 60 centers established in the United States (including Alaska and Hawaii) sufficient quantities of test booklets, answer sheets, questionnaires, and supervisor's manuals, to test the number of candidates registered to take the NBRT Written Examination - Therapist Level;

3. pay all shipping charges in connection with the shipment of test materials to the test centers and their return to ETS;

4. pay for all test center expenses including supervision honoraria;

5. ship the used answer sheets and biographical questionnaires to the Psychological Corporation in three batches on March 7, March 12, and when the final shipment is received from the remaining test centers;

6. as soon as all testing materials have been received from the test centers and accounted for, return to the Psychological Corporation, all used and unused test booklets, all unused answer sheets, manuals, and unused questionnaires;

7. initiate procedures to recover any test materials not returned to ETS from the test centers and to assist in the investigation of any incidents relating to the security of the Proficiency Test;

8. within two weeks after receipt of the punched and interpreted tab cards from the Psychological Corporation:

a. match the Proficiency Test total score and subscores to the Therapist Level test total score and subscores for each candidate; and

b. produce a tab card for each set of matched test scores (individual candidate identification will not be provided) and forward them to the Psychological Corporation;

9. Produce a roster (in descending therapist level score order) of all matched scores (individual candidate identification will not be provided); and

10. compute for each candidate taking the Proficiency Test the additional credit earned as directed by the NBRT Written Examination Committee.

ETS will endeavor to meet the schedule of commitments herein. However, it is agreed that this commitment is also contingent upon the timely discharge of responsibilities by AART, the Psychological Corporation, and NBRT.

As full and complete compensation for the ETS services provided under this agreement, AART agrees to pay to ETS the fixed sum of ~~$6,725.00~~ $5,000.00 upon completion of these services by ETS.

If the terms of this agreement are satisfactory, would you please return one original signature copy to me after both copies have been signed by AART.

Sincerely,

Russell W. Martin, Jr.
Assistant Treasurer

HRA PROJECT OFFICE
AMERICAN ASSOCIATION FOR
RESPIRATORY THERAPY

By William W. Johnson

Title _____

VW:RWM:h

A REVIEW OF METHODOLOGY FOR CRITERION-REFERENCED SURVEY TESTING*

The topic of criterion-referenced measurement has received considerable attention during the past decade. Much of the initial controversy that was generated over the relative merits of criterion-referenced and norm-referenced measurement appears to have subsided. Today, most psychometricians seemingly agree that criterion-referenced and norm-referenced measurement have different purposes, and that each is appropriate under the circumstances for which it was intended. Norm-referenced measures are generally more appropriate in selection situations while criterion-referenced instruments facilitate classification decisions regarding an examinee's position relative to a specified objective. The determining factor in the selection of a measurement technique is the type of information required by the decision maker.

More recently, attention has centered around the inadequacy of theoretical and methodological considerations for criterion-referenced tests. Many writers have discussed the inadequacies of classical measurement theory for criterion-referenced situations, and a few have attempted to modify existing theory to handle criterion-referenced cases. Still other writers have tried to develop new theories for criterion-referenced measurement. As new theories and methodologies emerge, they should be tested and evaluated to determine their viability under diverse circumstances.

One inhibitor of progress in the development of criterion-referenced test theory may be the widespread lack of uniformity in definition and use of basic terminology. A cursory examination of the criterion-referenced literature reveals considerable variation in the application of the concept. For example, investigators at one extreme emphasize the direct linkage between criterion-referenced test items and specified performance objectives. Kriewall (1972) advocated an item sampling approach in which "a learning objective (LO) is defined by a specified item population [p. 5]." An opposing philosophy was presented by Fremer (1972) who discussed criterion-referenced "interpretations" of tests and indicated that it is not necessary for tests to directly measure the criterion behavior; they must only lead to the same conclusions that result from sampling the behavior. Obviously, the characteristics of these two types of tests may be quite different. The item sampling approach employs item generation rules that insure direct correspondence between the items and the objective, thus producing intrinsically valid tests. In the case of tests providing criterion-referenced interpretations, no restrictions are placed upon the generation of items, and validity of the interpretations must be demonstrated by correlational or other validation procedures.

Recently, criterion-referenced measurement has been introduced into survey achievement testing situations such as state educational

*Contributed by Darol L. Graham, Ph.D., Assistant Professor, School of Allied Health Sciences, University of Texas Health Science Center, Dallas, Texas

assessment and proficiency testing programs. Many governing and credentialing bodies have adopted the assumption that precise measurement of educational objectives is essential for maintaining accountability. In establishing competency-base assessment programs, the objective-based measurement techniques that have proven so useful for making instructional development and management decisions provided an obvious tool. The logic of such an extension in the use of criterion-referenced instruments cannot be argued; however, the decision to employ such instruments was made without evidence of the suitability of criterion-referenced measurement for large-scale testing situations. It appears that the methodology for developing and evaluating criterion-referenced survey instruments should receive more attention before their use in large-scale assessment programs becomes widespread.

TEST CONSTRUCTION

A variety of procedures have been used in the development of criterion-referenced competency tests. Variation in test construction procedures are primarily related to the manner in which the objectives are used in the generation and refinement of test items. At one extreme are the instruments for which items are generated by randomly sampling from a precisely defined item universe. Such tests represent the most restricted form of the class of instruments known as "domain-referenced" tests. Domain-referenced tests are characterized by the use of domain-sampling procedures (Millman, 1973) for the generation of items. By utilizing such procedures for test construction, a high degree of content validity is ensured for the test. If the domain is sufficiently restricted to define a unitary skill, the items will form a homogeneous set that provides highly reliable measures.

In the development of domain-referenced tests, a number of procedures for defining an item universe and for generating test items have been suggested. One of the most straightforward and usable procedures involves the development of amplified objectives (Popham, 1972). Procedures have also been suggested by Osburn (1968), Hively, Patterson, and Page (1968), Kriewall (1969), Bormuth (1970), and others. Although all of these procedures seem appropriate for constructing criterion-referenced competency measures, not all of them utilize the item-sampling techniques required for generating a direct measure of the criterion in a manner that ensures content validity. By using a procedure such as Kriewall's (1969) that "provides an algorithm for test construction and thus secures the content validity required [p. 48]," it is possible to generate instruments with sufficient validity to serve as criterion measures in validation studies. In the remainder of the present paper, the term criterion measure will specifically refer to a criterion-referenced instrument constructed by such procedures.

Criterion-referenced tests at the other end of the continuum are constructed by a variety of procedures that do not depend on

the prior definition of an item domain. These tests can be considered criterion-referenced only to the extent that they provide "criterion-referenced interpretations" (Fremer, 1972). A test's ability to provide criterion-referenced interpretations generally must be demonstrated by correlational or other empirical methods. Thus, a criterion-referenced test of this type is not expected to actually "measure" a competency but only to provide data that correlates highly with the competency variable.

Specific procedures for developing the type of instrument described above have not been designated. Such instruments are often developed by simply identifying existing items that "appear" congruent with a specified objective. In defending the use of items that may not be directly related to an objective, Fremer (1972) stated, "A sample of tasks covering a number of objectives can permit sound inferences to whole classes of objectives, including many not represented in the sample [p.4]." Fremer indicated his motivation for such an approach by further stating, "The use of a survey test as a basis for making criterion-referenced inferences permits considerable efficiency in testing [p. 5]." Granted, the ability to use existing survey tests for making criterion-referenced interpretations would be highly advantageous for test publishers, empirical evidence of the validity of competency classifications provided by such instruments does not appear sufficient to warrant this approach at the present time.

A more typical approach, described by Popham (1970), is "to develop the test items with whatever generation rules are available, then try the items out to discover empirically which items are defective, that is, are not congruent with the criterion [p. 2]." This method was ployed by Hills (1970) in graduate level measurement and statistics classes. Other examples of the use of such procedures for the development and refinement of test items were provided by Iven (1970), Hsu (1971), and Olson (1974). The major problem with this approach is a general lack of proven empirical methods for determining the adequacy of items (Popham, 1970).

TEST VALIDITY

It is generally assumed that a domain-referenced test possesses high content validity by virtue of the judged adequacy of the domain definition and the procedures used to generate the test items. The validity of a criterion measure, however, is not dependent on any judgment or logical analysis of the item generation procedures. The item-sampling model presented by Kriewall (1972) started "with the assumption of a prima facia content validity [p. 4]." Kriewall explained that "The essential metaphor that enables one to meet this condition is the notion that a learning objective (LO) is defined by a specified item population [p. 4]."

Mosier (1947) took a similar stance in suggesting the concept of validity by definition. He further emphasized that "the direction

of the argument flows from the test to the definition of the criterion rather than from the conceptually defined criterion to the test as a valid measure [p. 196]." Additionally, he pointed out that the only proper statement permitted by definitional validity is "This test is a valid measure of that and only that universe of individual behavior patterns for which these items constitute a representative sample [p. 196]." Finally, Mosier stated, "The objective of the test is so defined that the index of reliability (the square root of the reliability coefficient) is, by definition, the measure of validity [p. 192]." Jackson (1970) also described definitional validity and suggested that in such situations reliability "is considered a sufficient, rather than only a necessary condition for validity [p. 13]."

Fremer (1972) proposed the use of a number of procedures for validating tests that purport to provide criterion-referenced interpretations. Whenever it is possible to carefully construct a criterion measure, the most appropriate expression of validity would probably be a correlation coefficient indicating the concurrent validity (Cronbach, 1970) of the competency predictor. Ebel (1961) discussed many of the problems associated with the selection or development of adequate criterion measures for validating a test and indicated that considerable effort must go into the establishment of a suitable criterion.

In practice, it is frequently assumed that, by using the objective as a guide, a person who is thoroughly familiar with the content domain can produce reasonable valid items. A further indication of the content validity of such tests is often provided by the judgment of additional content specialists. Seldom does the test developer employ empirical procedures to examine or refine the item set to demonstrate concurrent validity with a more direct measure of the criterion.

It is believed that one of the most serious mistakes made by test developers is to follow procedures similar to those outlined above for generating criterion-referenced tests without obtaining empirical verification of test validity. Graham (1974) presented evidence that the assumption of a content expert's ability to accurately judge the validity of a test should be seriously questioned. Although some experts may be able to judge the validity of some tests, exclusive reliance upon such expedient procedures should be discouraged. Certainly, the validation of important tests to be used for assessing the competency of individuals should not be limited to the opinions of judges, regardless of their qualifications.

TEST RELIABILITY

Due to many of the initial uses of criterion-referenced tests, it was conceivable, and sometimes considered desirable, to anticipate

virtually zero true score variance in a single administration of a test. Popham and Husek (1969) were correct in noting that variability is irrelevant and not an essential condition for a good criterion-referenced test. In reality, however, such a situation only exists when a restriction test sample is used. Whenever a criterion-referenced test is administered simultaneously to members of both the mastery and nonmastery populations, considerable score variability is expected and found.

In speaking of tests generally, Cronbach (1951) made it clear that "in a homogeneous test, the items measure the same things [p. 154]" and that "if a test has substantial internal consistency, it is psychologically interpretable [p. 154]." In a discussion of reliability, Stanley (1971) demonstrated that the only time dichotomously scored items can be perfectly intercorrelated, resulting in a maximum value of one for KR-20, is when all items have equal difficulty. Since the goal in criterion measure construction is to develop a homogeneous set of items that all measure the same intellectual skill to the same extent, such an indicator of test homogeneity appears to be an appropriate expression of criterion measure realiability.

Criterion-referenced tests which are constructed by other procedures would not be expected to contain items with the same degree of consistency as a test with items randomly sampled from a homogeneous domain. Nevertheless, internal consistency would also seem essential for a reliable instrument of this type and KR20 values should be considered.

ITEM CHARACTERISTICS

Item Discrimination

A number of different discrimination indices have been proposed for criterion-referenced tests. Many of the procedures for computing the various indices are based upon the assumption that a valid criterion-referenced test should be sensitive to instruction. These procedures generally reflect the original work of Cox and Vargas (1966) involving differences in pre- and posttest performance. Hsu (1971) commented on the fact that instruction and learning are not necessarily equivalent and suggested that discrimination between individuals in the learned and unlearned states might be more appropriate. The method proposed by Hsu involved the determination of membership in the mastery and nonmastery populations according to a predetermined cut-off score on a single administration of the test. Such an index (D_p) of the proportions of correct responses between mastery and nonmastery groups seems appropriate for describing the characteristics of criterion-referenced test items.

Another frequently used item discrimination index is the item-test correlation coefficient (ϕ). In computing the ϕ index, Hsu (1971) employed the procedure described above for determining membership in the mastery and nonmastery populations. This procedure

of assigning mastery according to a predetermined cutoff score on a single administration of a test appears useful for computing ϕ coefficients for the items of criterion-referenced competency tests.

Although the information provided by the two item discrimination indices is somewhat redundant, phi provides a modification of the difference index based upon the relative item difficulties. The relationship between ϕ and D_p is:

$$\phi = D_p \sqrt{\frac{p_t q_t}{p_i q_i}}$$

where p_t and q_t represent the respective proportions of masters and nonmasters on the test and p_i and q_i represent the respective proportions of examinees that responded correctly and incorrectly to the item under consideration. Thus, p_t and q_t provide an indication of test difficulty while p_i and q_i express the item difficulty. Further investigation appears necessary to determine whether the two indices are sufficiently different to warrant the determination of both or if one is to be preferred over the other.

Item Difficulty

Item difficulty (p) is defined as the proportion of examinees that answer a given item correctly. For a criterion-referenced test, item difficulty is sometimes considered a function of the learning state of the examinee. Thus, items should be uniformly difficult for nonmasters of an intellectual skill and uniformly easy for masters of the skill. Whenever an examination sample is comprised of both masters and nonmasters of an intellectual skill, the magnitude of the p value for an item depends upon the relative representation of the two mastery populations. Since the items of a criterion measure are randomly sampled from the same precisely defined domain, all items in the measures should have similar difficulties for a given group of examinees. Thus, homogeneity of item difficulty would be considered a desirable characteristic for a criterion measure.

In addition to examining the p values calculated from performance by the entire test sample, Graham (1974) considered the item difficulty values separately for responses arising from examinees assigned mastery and nonmastery status by a test. Hypothetically, item difficulty values for mastery and nonmastery populations should be one and zero, respectively, on a homogeneous test. Items that deviate from either of these ideal values would be systematically biased for one or the other of the competency populations. This useful information is lost when the difference between these two item difficulty values is calculated in the determination of the item discrimination index, D_p.

TEST LENGTH AND PASSING SCORES

The procedures for determining the number of items and cutting scores needed for criterion-referenced tests are attracting increasing

amounts of attention. Emrick (1971) presented a model for determining cutting scores based upon the assumption that mastery of a unitary skill is demonstrated as a binary variable. He further assumed, however, that mastery can occur in degrees because of the simultaneous measurement of a collection of skills. The model for assigning mastery status was based upon a costbenefit analysis of the consequences of different types of measurement error. Although the model may be useful for certain types of criterion-referenced tests, the procedures for estimating costs associated with the various types of measurement errors have not been well formulated.

Other methodologies involving decision rules based upon acceptance sampling models (Kriewall, 1969), classical binomial models (Millman, 1973) and Bayesian models (Novick, Lewis, & Jackson, 1973) have been proposed. These methodologies all appear relevant for situations involving continuous variables such as scores on proficiency tests with limited homogeneity. These approaches do not appear necessary for criterion measures but may be of some value for considering measures that are designed to give criterion-referenced interpretations.

Several factors that influence the number of items needed for a criterion-referenced test were identified by Graham (1974). These factors were: (1) intended use of the test, (2) item homogeneity, (3) response format, and (4) relative seriousness of different types of classification errors. It was concluded that for important tests of unitary intellectual skills comprised of free-response items in which false positives are at least as important as false negatives, the minimum test length for reasonably reliable classifications may be four or five items. As any of these factors change, the number of items needed for reliable measures will correspondingly increase. Specific guidelines for determining the length of various types of criterion-referenced tests skills should be developed.

SUMMARY AND CONCLUSIONS

The most difficult task facing the developer of criterion-referenced tests is the statement and operationalization of performance objectives. Failure to clearly delineate and sample from the domain of behaviors that define an objective precludes the establishment of definitional validity for a measure. Once the item generation rules are specified, however, test construction becomes a routine operation. As long as items are randomly generated according to the rules specified, the items are representative of the behaviors which comprise the domain.

It is believed that the overall quality of criterion-referenced instruments currently used for survey testing, whether for groups or individuals, could be improved considerably. Procedures are emerging which may provide a means of enhancing the quality of many existing survey tests. Admittedly, such procedures require extensive, painstaking attention to the identification or development of a

suitable criterion. Criterion-referenced interpretations derived from the type of instruments typically used for survey testing may not be justified, however, until the relation of the test to the objective has been adequately demonstrated through empirical validation procedures.

Many of the inadequacies of criterion-referenced tests can be traced to the intended use of an instrument. In large-scale survey testing situations, efficiency of administration is, by necessity, one of the primary considerations. The volume of data collected generally dictates that the items used in such situations be in multiple-choice or similar format to facilitate scoring. To enable sufficient coverage, the objectives must often be rather global and the number of items per objective restricted to three or four, sometimes less. Since reliability is a necessary condition for test validity, each of these factors creates problems for criterion-referenced survey testing.

The use of multiple-choice items introduces a substantial amount of random error into item data. Normally, the effects of guessing are not considered critical because of adequate test length. For a test of three or four items, however, the effects can be appreciable, and a number of examinees can be misclassified by chance. In addition, the testing of somewhat global objectives tends to reduce the homogeneity of items, further decreasing test reliability. In some instances, it may be possible to analyze intellectual skill domains to identify hierarchical dependencies that would permit some sort of sequential or convergent testing strategy. In general, however, the task of constructing short, multiple-choice tests over objectives of the desired breadth, with sufficient reliability to make valid competency decisions about individuals, is extremely difficult.

In conclusion, the use of criterion-referenced instruments for survey testing greatly increases the need for adequate theories and methodologies relating to criterion-referenced measurement. In classroom management situations, test quality is seldom critical. Other information sources provide a constant check on the criterion-referenced data. Since instructional management is a continuously ongoing process and most classroom decisions are of a temporary nature, decisions based upon invalid or inaccurate data can be readily modified at any time. On the other hand, survey testing often represents a single data collection effort and constitutes the sole information source for the decision maker. If the results of such testing are likely to have far-reaching effects upon the examinees, the integrity of the data is critical.

REFERENCES

Adkins, L.M., Lipner, L.J., Maxson, B.M., and Graham, D.L. Clinical teacher desired pupil behaviors: An individualized mathematics curriculum. Tallahassee: Clinical Teacher Model, Florida State University, 1973.

Bormuth, J.R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Brennan, R.L. Some statistical problems in the evaluation of self-instructional programs. Research Memorandum No. 1, June, 1970, Harvard University, Cambridge, Massachusetts.

Cox, R.C. and Vargas, J.S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, February, 1966.

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L.J. Essentials of psychological testing. (3rd ed.) New York: Harper & Row, 1970.

Davis, F.B. 1971 AERA Conference Summarys: II. Criterion Referenced Measurement. TM Reports No. 12, March, 1972, ERIC Clearinghouse on Tests, Measurement, & Evaluation, Educational Testing Service, Princeton, New Jersey.

Ebel, R.L. Must all test be valid? American Psychologist, 1961, 16, 640-647.

Ebel, R.L. Content standard scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Emrick, J.A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 4, 321-326.

Florida State-Wide Eighth-Grade Testing Program. Technical Report No. 1-73, January, 1973, College of Education, Florida State University, Tallahassee, Florida.

Fremer, J. Criterion-referenced interpretations of survey achievement tests. Test Development Memorandum 72-1, January, 1972, Educational Testing Service, Princeton, New Jersey.

Gagne, R.M. The acquisition of knowledge. Psychological Review, 1962, 69, 355-365.

Gagne, R.M. Instructional variables and learning outcomes. CSEIP Occasional Report No. 16, September, 1968, University of California, Los Angeles, California.

Gagne, R.M. Learning Lierarchies. Educational Psychologist, 1969, 6, 2-6.

Gagne, R.M. The conditions of learning. (2nd Ed.) New York: Holt, Rinehart, & Winston, 1970.

Gagne, R.M. Domains of learning. Presidential address presented at the annual meeting of the American Educational Research Association, February, 1971, New York City, New York.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, R. A criterion-referenced test. In W.J. Popham (Ed.), Criterion-referenced measurement: An introduction. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Graham, D.L. An empirical investigation of the application of criterion-referenced measurement to survey achievement testing. Unpublished doctoral dissertation, Florida State University, 1974.

Hambleton, R.K., & Novick, M.R. Toward an integration of theory and method for criterion referenced tests. Journal of Educational Measurement, 1973, 3, 159-170.

Hartnett, R.T. Accountability in higher education: A consideration of some of the problems of assessing college impacts. Princeton, N.J.: College Entrance Examination Board, 1971.

Hills, J.R. Experience in small graduate classes and approaches to evaluating criterion-related measure. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, March, 1970.

Hively, W. Domain-referenced testing. Unpublished paper, Minne-mast Project, University of Minnesota, 1972.

Hively, W., Patterson, H.L., & Page, S.H. A "universal-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Hsu, T.C. Empirical data on criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York City, February, 1971.

Ivens, S.H. An investigation of item analysis, reliability and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, 1970.

Jackson, R. Developing criterion-referenced tests. TM Report No. 1, June, 1970, ERIC Clearinghouse on Tests, Measurement, & Evaluation, Educational Testing Service, Princeton, New Jersey.

Kriewall, T.E. Application of information theory and acceptance sampling principles to the management of mathematics instruction. Technical Report No. 103, October, 1969, Wisconsin Research and Development Center, Madison, Wisconsin.

Kriewall, T.E. Aspects and applications of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1972.

Mayo, S.T. Mastery learning and mastery testing. NCME Measurement in Education, 1970, 1, (3).

Miller, R.G. Simultaneous statistical inference. New York: McGraw-Hill, 1966.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-215.

Mosier, C.I. A critical examination of the concepts of face validity. Educational and psychological measurement, 1947, 7, 191-205.

Novick, M.R., Lewis, C., & Jackson, P.H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-46.

Olson, M.A. A comparison of three techniques for selecting items for criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, 1974.

Osburn, H.G. Item sampling for achievement testing. Educational and Psychological Measurement, 1969, 28, 95-104.

Popham, W.J. Indices of adequacy for criterion-referenced test items. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, March, 1970.

Popham, W.J. Selecting objectives and generating test items for objectives-based tests. Los Angeles: Instructional Objectives Exchange, 1972.

Popham, W.J., & Husek, T.R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 1, 1-9.

Stanley, J.C. Reliability. In R.L. Thorndike (Ed.), Educational measurement. (2nd Ed.) Washington: American Council on Education, 1971.